Entity Consolidation and Alignment in Semi-Structured Data Sources

Eduard Hovy, Andrew Philpot and Patrick Pantel
Information Sciences Institute
University of Southern California
4676 Admiralty Way
Marina del Rey, CA 90292
{hovy,philpot,pantel}@isi.edu

ABSTRACT
A large portion of collected data is stored in semi-structured form, i.e., organized or related through columns or lists but without any formal schema or metadata. For example, many government organizations and companies feature employee directories and project affiliations in HTML tables on their websites. Making sense of these requires automatic methods for data alignment, matching and/or merging. Here, we describe *Guspin*, a tool for automatically consolidating entities and for aligning data across semi-structured data sources. Our project, based on principles of information theory, measures the relative importance of data, leveraging them to quantify the similarity between entities. We have applied our technology to discover duplicates and perform alignments for data sources provided by the Environmental Protection Agency and California environmental agencies.

Categories and Subject Descriptors
H.3.3 [**Information Storage and Retrieval**]: Information Filtering.

General Terms
Algorithms, Experimentation.

Keywords
Information theory, mutual information, semi-structured data, equivalence class detection, entity consolidation.

1. *Guspin*[1]: A Data Modeling Portal
Guspin is a general purpose tool for finding equivalence classes, consolidating entities, and aligning data within a population of semi-structured data. It provides a simple user interface where a user uploads one or multiple data files containing observations for a population. The system inspects the data, identifies patterns, suggests alignments and consolidations, provides a browsing interface for viewing the analysis, and permits downloading of the analysis for further examination.

\

[1] *Guspin* is available from http://guspin.isi.edu.

2. Case Study A: Entity Consolidation
The Environmental Protection Agency (EPA) maintains a centrally managed database, called the Facilities Registry System (FRS), recording American facilities subject to environmental regulations (e.g., refineries, gas stations, manufacturing sites, etc.) Duplicates exist in the FRS since it is compiled from various local and state jurisdictions, which often have different ways of representing data. Our goal on this data set is to automatically discover the duplicate entries.

We obtained from the EPA a sample of the FRS. Each record includes the address, state, zip code, facility name, etc. for a particular facility. Through *Guspin*'s web interface, we upload the FRS data and then *Guspin* measures the mutual information between entities and observations (e.g., address, emission statistics, codes, etc.), computes the similarity between each pair of entities, and clusters entities into equivalence classes. One can search for individual entities by using *Guspin*'s search feature. For example, *Guspin* discovered that facility 189 consolidates with facilities 300 and 79. Figure 1 shows the results of launching a search for facility 189's most similar entities (i.e., potential duplicates). For each similar entity, the similarity score is shown along with a "*why?*" link, which enables the user to compare the observations of the two facilities (important observations are used to compute the similarity between entities).

Figure 2 illustrates two such comparisons: a) a comparison between the observations for facilities 189 and 79; and b) a comparison between the observations for facilities 189 and 300. Observations colored in blue and in green were observed for only one of the two facilities. Red observations, however, were shared by both facilities. Figure 2 lists observations in descending order of mutual information scores. For very similar entities, we therefore expect that most important observations (those at the top of the list) will be colored red. In fact, note that even though Figure 2 shows that facilities 189 and 79 share fewer common observations than facilities 79 and 300, the similarity between facilities 189 and 79 is greater since more *important* features are shared (i.e., they have more red features at the top of the list).

3. Case Study B: Data Alignment
Continuing our relationship with the California Air Resources Board (CARB) and various California Air Quality Management Districts (AQMDs), built during our DG project, we obtained emissions inventories (a comprehensive description of emitters and emission statistics) submitted annually by California AQMDs to CARB. We applied *Guspin* to the inventories to automatically

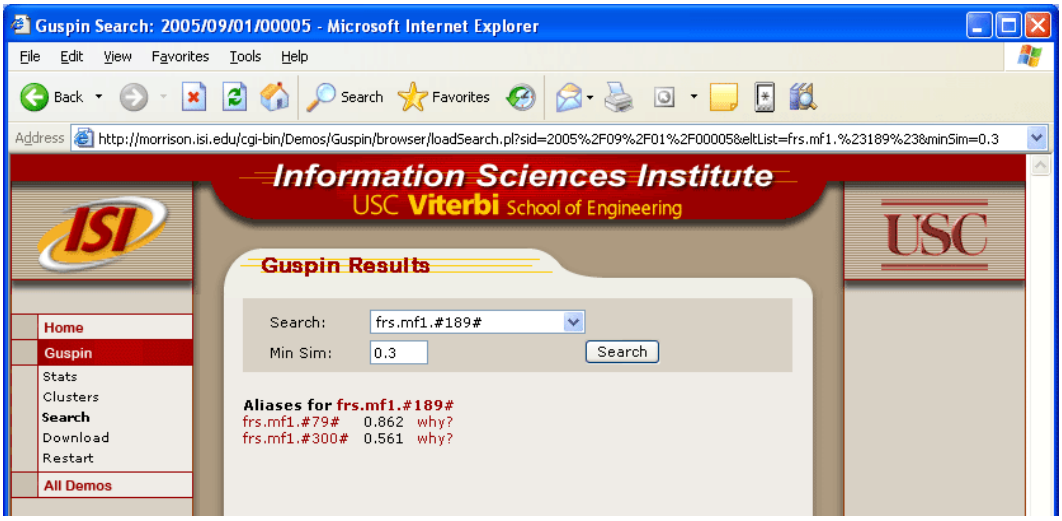

Figure 1. *Guspin*'s search interface for displaying an entity's most similar entities. In this example, we see that facility 189 from EPA's Facilities Registry System is most similar to facilities 79 and 300. Clicking on a facility displays its observations. Clicking on "*why?*" compares the observation data from facility 189 with those from facilities 79 and 300.

a)

b)

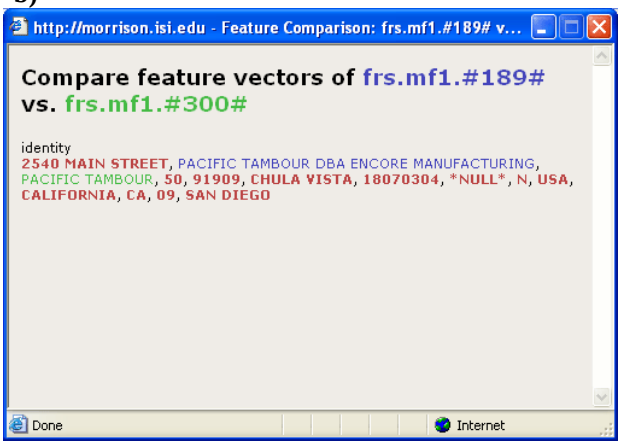

Figure 2. *Guspin* comparison of two entities' observations: a) comparison of the observations for facilities 189 and 79 (similarity = 0.862); b) comparison of the observations for facilities 189 and 300 (similarity = 0.561). Observations are sorted in decreasing order of pointwise mutual information scores. Observations colored in blue and green are shared by only one of two facilities, whereas red observations are shared by both.

discover overlapping entries across data sources. Below, we summarize *Guspin*'s performance on the CARB and Santa Barbara County Air Pollution Control District 2001 emissions inventories:

- with 100% accuracy, *Guspin* extracted 50% of the matching facilities;
- with 90% accuracy, *Guspin* extracted 75% of the matching facilities;
- for a given facility and the top-5 mappings returned by *Guspin*, with 92% accuracy, *Guspin* extracted 89% of the matching facilities.

4. Conclusions

Researchers, organizations and government agencies working with semi-structured data are in need of a tool for discovering duplicate or overlapping data. Our project, based on principles of information theory, measures the importance of observations and then leverages these to quantify the similarity between entities. Though the technology is applicable to a wide range of applications, we have built *Guspin* to address the general problems of consolidating entities and aligning data across semi-structured data sources. *Guspin* has been applied to solve real problems faced by the Environmental Protection Agency.

Guspin may be applied to several other tasks. For example, it can be used to identify occurrences of plagiarism in essays represented by the words they contain or it can be used to find co-regulated genes represented by their expressions in a series of micro-array experiments. It is our intention to promote *Guspin* to all government agencies as a portal for helping in the collection, disambiguation, and analysis of their data.